AD_____

Award Number: W81XWH-13-1-0335

TITLE: - Whole-Genome Sequencing of High-Risk Families to Identify New Mutational Mechanisms of Breast Cancer Predisposition

INITIATING PRINCIPAL INVESTIGATOR: Tomas Walsh, PhD

CONTRACTING ORGANIZATION: University of Washington, Seattle, WA, 98195

REPORT DATE: October 2014

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

Approved for public release; distribution unlimited

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*



| **1. REPORT DATE** *(DD-MM-YYYY)* | **2. REPORT TYPE** | **3. DATES COVERED** *(From - To)* |
|---|---|---|
| October 2014 | Annual | 15 Sep 2013 - 14 Sep 2014 |

**4. TITLE AND SUBTITLE**

Whole Genome Sequencing of High-Risk Families to Identify New Mutational Mechanisms of Breast Cancer Predisposition

**5a. CONTRACT NUMBER**

**5b. GRANT NUMBER**
W81XWH-13-1-0335

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**

Tomas Walsh, PhD, Mary-Claire King, PhD

email: twalsh@u.washington.edu

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**

University of Washington
Seattle, WA, 98195

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

**10. SPONSOR/MONITOR'S ACRONYM(S)**

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

**Approved for public release; distribution unlimited**

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**

As genes for inherited disease are increasingly well characterized by next generation sequencing approaches, it is clear that some mutations may act through promoters, enhancers, and other non-coding regulatory regions. Our hypothesis for this proposal is that much of the substantial remaining familial risk of breast cancer is due to a large number of individually rare alleles of moderate-to-severe effect located in the non-coding regions of the genome. For this proposal we will evaluate 30 large, extended kindreds severely affected with breast cancer, each of whom has been comprehensively evaluated in our lab by targeted genomic sequencing for mutations of all classes in all known breast cancer genes and by whole exome sequencing for coding region mutations exome-wide. These families are a unique discovery series for identification of regulatory mutations that may reveal new mutational mechanisms for breast cancer predisposition.

**15. SUBJECT TERMS**
Genome sequencing, mutation, breast cancer cancer susceptibility genes

| **16. SECURITY CLASSIFICATION OF:** | | | **17. LIMITATION OF ABSTRACT** | **18. NUMBER OF PAGES** | **19a. NAME OF RESPONSIBLE PERSON** |
|---|---|---|---|---|---|
| | | | | | USAMRMC |
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | 139 words | 7 | **19b. TELEPHONE NUMBER** *(include area code)* |

**Table of Contents**

**INTRODUCTION:** Despite tremendous advances in mutation detection with gene panels and exome sequencing the majority of high risk breast cancer families do not have their causative alleles identified from the protein-coding region of the genome. We hypothesize that their critical mutations lie in unknown regulatory regions of the genome. Through whole genome sequence analysis of severely affected families and functional annotation and experimental evidence we plan to identify new mutational mechanisms that predispose to breast cancer. Our ultimate goal is to enable information on newly identified mutations and mutational mechanisms to be useful to clinicians and to women and their families.

**KEYWORDS:** Breast cancer, *BRCA1*, *BRCA2*, whole genome sequencing, promoter, enhancer, transcription factor binding site, gene regulation, mutation.

**OVERALL PROJECT SUMMARY:** We describe below our progress during Year 1 with respect to each of the Tasks outlined in the approved SOW. All Tasks are performed at the same research location and the joint responsibility of both the Initiating PI (Tom Walsh PhD) and Partnering PI (Mary-Claire King, PhD).

*TASK 1. Perform whole genome sequencing of germline DNA from 100 breast cancer patients selected from 30 severely affected families.*

*1a. Prepare 100 standard paired end library with 300-400bp inserts (months 1-3)*
*1b. Prepare 100 mate-paired library with tightly defined 6kb inserts (months 1-3)*
*1c. Sequence the paired end and mate-paired libraries on a HiSeq2500 (months 2-9)*

We have prepared both library types and generated sequencing data on the 100 breast cancer patients.

*TASK 2. Annotating sequencing genome variants with respect to population frequency and overlap with ENCODE regions.*

*2a. Align reads to the reference sequence (months 4-10)*
*2b. Identify SNPs, indels, CNVs and rearrangements by bioinformatic tools (months 4-10)*
*2c. Filter variants from Task 2b against publically available databases to remove common events (months4-10)*
*2d. Filter rare and private variants from Task 2c within families to obtain segregating variants (months 4-10)*
*2e. Compare surviving events from Task 2d to ENCODE regions (months 4-10)*
*2f. Further filter variants from Task 2e to ENCODE variants mapped only in breast tissues/lines (months4-10)*

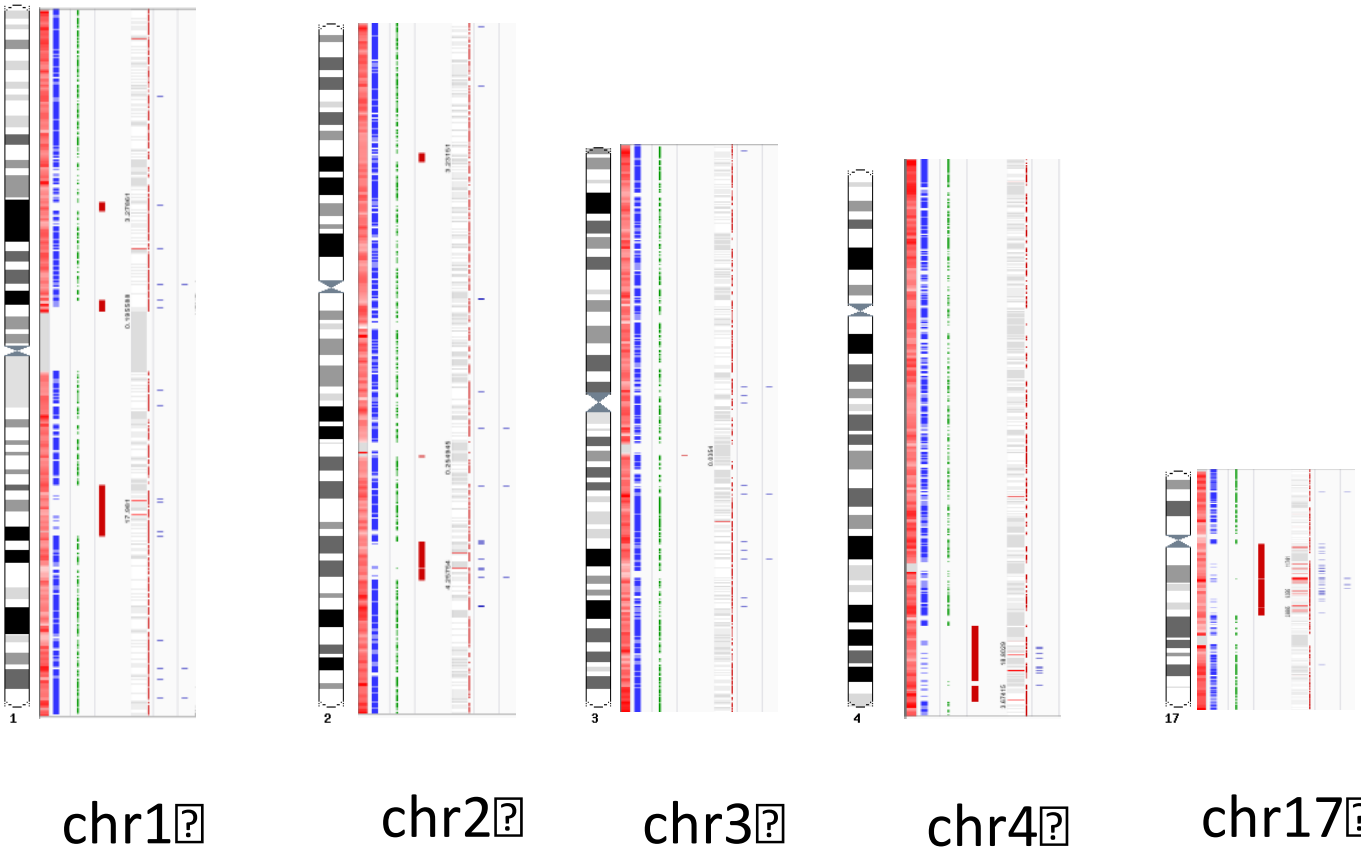
We have developed a functional annotation approach to filter variants from the whole genome sequences. Table 1 below summarizes the variant output from one of the severely affected breast cancer families.

**Table 1. Distribution of different types of mutations at different filtering levels in the whole genome sequencing data of two patients from a severely affected breast cancer Family 1041.**

| | All | Shared | Rare | Excluding IBD0 |
|---|---|---|---|---|
| Intergenic | 3,345,727 | 1,650,045 | 35,927 | 3,990 |
| ncRNA | 266,300 | 130,836 | 3,104 | 329 |
| Up-downstream | 72,754 | 35,055 | 865 | 98 |
| Untranslated | 48,884 | 23,492 | 435 | 49 |
| Intronic | 1,986,665 | 970,436 | 20,088 | 2,139 |
| Missense | 13,933 | 6,891 | 48 | 11 |
| Silent | 15,150 | 7,567 | 30 | 5 |
| Nonsense | 116 | 45 | 1 | 0 |
| Splice | 198 | 120 | 1 | 0 |
| Stoploss | 15 | 8 | 0 | 0 |
| In-frame | 385 | 162 | 2 | 0 |
| Frame-shift | 199 | 92 | 2 | 1 |
| Total | 5,750,326 | 2,824,749 | 60,503 | 6,622 |

We have filtered variants from the breast cancer families to those in the 1000Genomes project[1] and more recently to those described in the Genomes of the Netherlands[2]. We further narrowed down the variant list by filtering out non-shared segments of the genome (termed IBD0) in each of the 30 families. Figure 1 shows the non-IBD0 regions in Family 1041.

**Figure 1. Non-IBD0 regions for Family 1041. The largest region overlaps *BRCA1* on chromosome 17.**

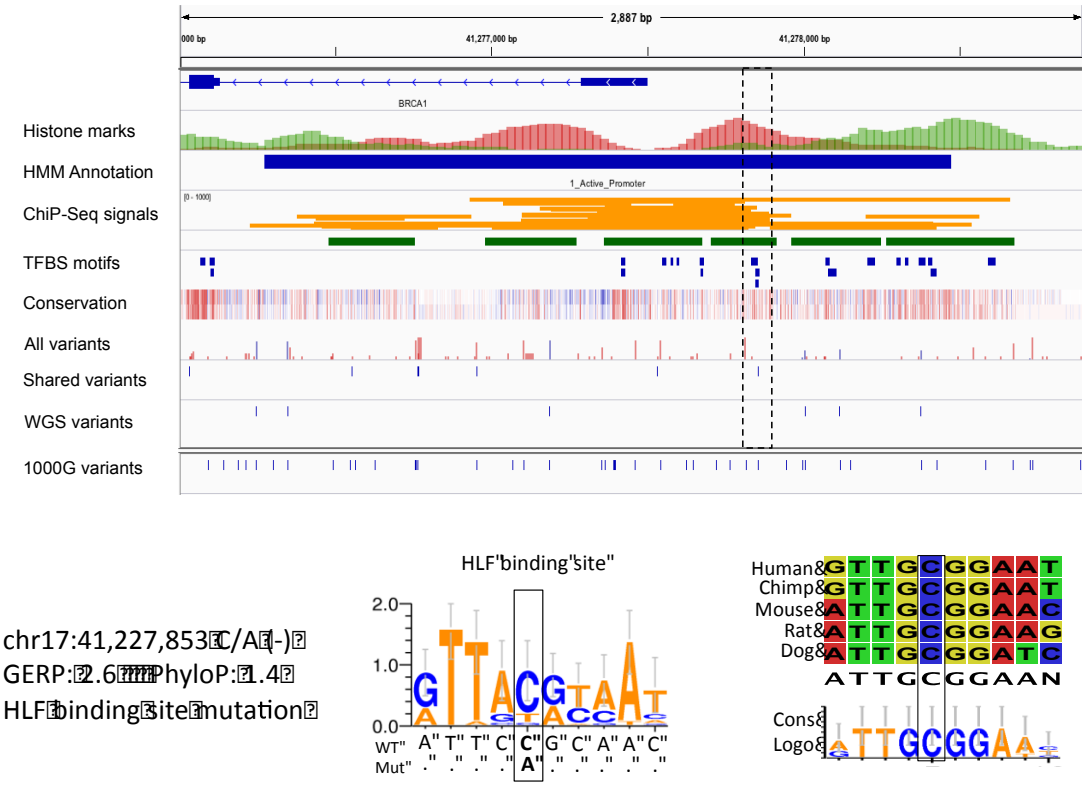



After the non-IBD0 sharing constraint has been applied we categorized the remaining variants with respect to the following features developed through the various projects of ENCODE[3].

1) Genomic location: gene upstream, 5'UTR, 1st exon and intron
2) Histone marks such as H3K9Ac, H3K27Ac)
3) DNAseI hypersensitivity data
4) ChIP-seq signals
5) Conservation of the variant with its flanking bases
6) Effect of variant on Transcription Factor binding site motif score using position weight matrices.

We show in Figure 2 an example of a variant from breast cancer Family 1041 that was shared by all women with breast cancer in the family and scored highly in our filtering scheme.

**Figure 2. A C>A variant located at chr17:41,227,852 identified in Family 1041. The variant is located within a potential regulatory region upstream of *BRCA1* and alters the fifth base of the conserved HLF binding site. The region is conserved in 5 mammalian species.**




*TASK 3. Characterize potential regulatory variants.*

*3a. Generate enhancer constructs with wildtype and variant regulatory regions (months 8-16)*
*3b. Transfect constructs into cell lines, monitor luciferase activity (months 8-16)*
*3c. Measure gene expression in patients' lymphoblasts (months 8-16)*

We have made the mutant and wildtype constructs of 11 different potential non coding regulatory mutations including chr17: 41,227,852 C>A from Family 1041 and are currently assessing luciferase activities. Direct mRNA measurements are also ongoing for these variants at their associated genes in patient's lymphoblasts.

*TASK 4. Resequence mutant regulatory regions in large series of patients to identify additional mutations*

*4a. Design molecular inversion probes (MIPs) for promising regulatory regions (months 12-24)*
*4b. Perform MIP amplification, hybridization and sequencing (months 12-24)*
*4c. Annotate varints within regulatory regions with respect to frequency (months 12-24)*
*4d. Statistical analysis of variants (months 12-24)*

We have begun generating candidate lists for resequencing and will commence Task 4 shortly.

**KEY RESEARCH ACCOMPLISHMENTS:**

- Developing an approach that functionally annotates whole genome sequencing data with respect to shared, rare and potential regulatory variants is a major contribution to achieving our goals.

**CONCLUSION:** In Year 1 of this project we have generated all the whole genome sequencing data necessary to achieve our goal of identifying new mutational mechanisms for breast cancer predisposition. We have developed a functional annotation approach that can pinpoint potential regulatory mutations and we are beginning to develop a list of variants that will be furthered evaluated experimentally.

Our objective of identifying new mutational mechanisms of breast cancer predisposition is on track and can be achieved with further analysis and the experiments outlined in our SOW Tasks.

**PUBLICATIONS, ABSTRACTS, AND PRESENTATIONS:** At the end of Year 1 we have not submitted any manuscripts but anticipate doing so in Year 2 and presenting our findings at scientific meetings in 2015.

**INVENTIONS, PATENTS AND LICENSES:** Nothing to report

**REPORTABLE OUTCOMES:** Nothing to report, at this stage

**OTHER ACHIEVEMENTS:** Nothing to report

**REFERENCES:**

1. 1000 Genomes Project Consortium, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012 Nov 1;491(7422):56-65. doi: 10.1038/nature11632.

2. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014 Aug;46(8):818-25. doi: 10.1038/ng.3021.

3. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, Dunham I, Elnitski LL, Farnham PJ, Feingold EA, Gerstein M, Giddings MC, Gilbert DM, Gingeras TR, Green ED, Guigo R, Hubbard T, Kent J, Lieb JD, Myers RM, Pazin MJ, Ren B, Stamatoyannopoulos JA, Weng Z, White KP, Hardison RC. Defining functional DNA elements in the human genome. Proc Natl Acad Sci U S A. 2014 Apr 29;111(17):6131-8. doi: 10.1073/pnas.1318948111.

**APPENDICES:** None